AD________________

Award Number: W81XWH-11-2-0133

TITLE: Framework for Smart Electronic Health Record-Linked Predictive Models to Optimize Care for Complex Digestive Diseases

PRINCIPAL INVESTIGATOR: Michael A. Dunn, MD

CONTRACTING ORGANIZATION: The University of Pittsburgh

Pittsburgh, PA 15213-3320

REPORT DATE: June 2012

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE June 2012	**2. REPORT TYPE** Annual	**3. DATES COVERED** May 12, 2011 – May 11, 2012

4. TITLE AND SUBTITLE

Framework for Smart Electronic Health Record-Linked Predictive Models to Optimize Care for Complex Digestive Diseases

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH11-2-0133

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Michael Dunn, MD, Melissa Saul, MS

E-Mail: dunnma@upmc.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The University of Pittsburgh
Pittsburgh, PA 15213-3320

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Our major objective is to develop an electronic application capable of integrating and semantically standardizing electronic medical record (EMR) data to generate de-identified datasets populated with longitudinal clinical data drawn from diverse sources. In Year 1 of our project, we have successfully built the infrastructure to support this project. We have defined and generated the EMR-based datasets to be used for algorithm development.

In year 2, we will test the ability of this tool to support predictive modeling of the outcomes of complex digestive diseases using Bayesian network (BN) analysis of the generated databases. We will further compare performance among models generated using EMR data alone and data from disease-specific clinical research repositories (with and without genetic data). In collaboration with Walter Reed Army Medical Center, we will share our data acquisition strategies and algorithmic model development. The integration of the two distinct patient populations will lay the groundwork for future data-sharing projects of mutual interest.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	11	**19b. TELEPHONE NUMBER** *(include area code)*

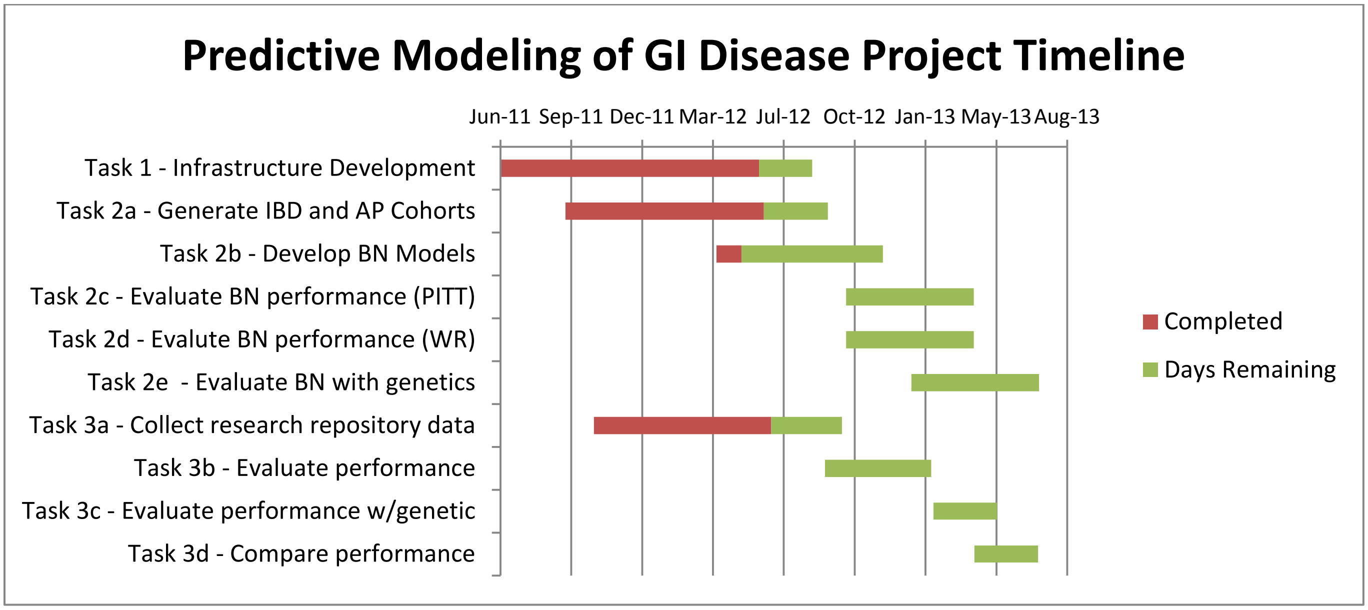
Contents

Introduction

Complex disorders result from the interaction of genetic, metabolic, and environmental factors that may not by themselves produce disease but can combine to alter disease severity and its progression. These factors, which may be contained in an electronic medical record (EMR) system, can be used to build predictive models of disease with the hope of improving disease management.

It is difficult to find these factors in EMR systems as the information is in both structured and unstructured formats that have been collected over many years. Research studies, in contrast, only collect a limited snapshot of a patient's clinical history. This information is usually not rich enough to develop predictive models. To construct a useful patient profile for analysis requires collecting disease progression and treatment information from a wide variety of sources that may span twenty years or more.

Our study goal is to develop the Megascope application to provide a software platform for the integration of clinical, genomic and research data collected from multiple sources. The University of Pittsburgh's Department of Biomedical Informatics (DBMI) and Division of Gastroenterology is an ideal collaboration to achieve this goal given our history of successful development of informatics applications and clinical research in complex GI diseases.
We will test the ability of Megascope to support predictive modeling of the outcomes of complex digestive diseases using Bayesian network (BN) analysis of the generated databases. We will further compare performance among models generated using EMR data alone and data from disease-specific clinical research repositories (with and without genetic data).



Body

Progress report on Technical Objective 1 – Infrastructure Development

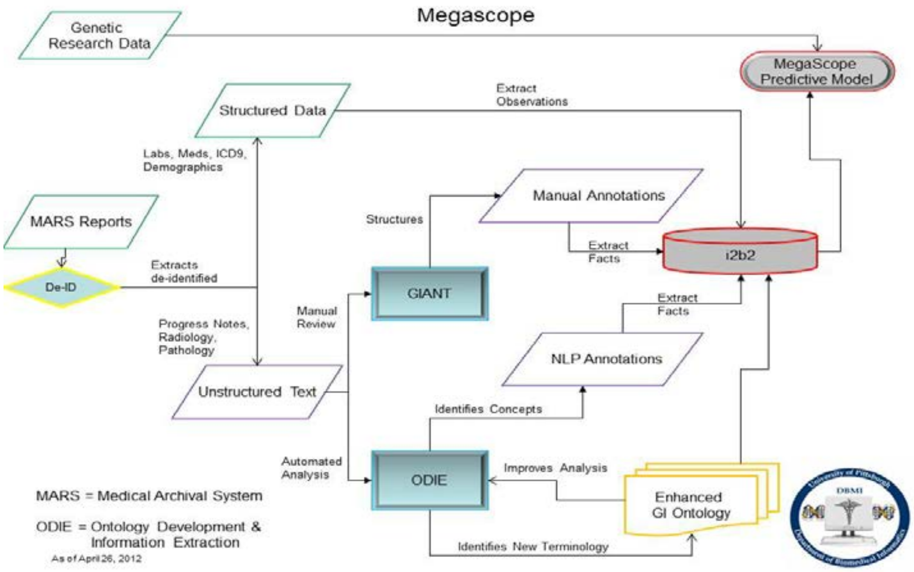
During the past year we have successfully built an application called Megascope to support our project goals. We realized early in the project that we needed to have a robust, open-source platform that would support the integration of clinical and genetic data. We also wanted to have an application that would not require a lengthy development cycle for creating data models. We decided to use the i2b2 (www.i2b2.org) framework to aggregate various sources of clinical and genomic data into a common vocabulary. This conversion to a common vocabulary, technically referred to as a controlled vocabulary or ontology, allows for us to treat many sources of data as though they are one. The i2b2 structure also has support for data mappings using LOINC and RxNORM enabling the data to be stored in uniform nomenclature. We plan to use various existing components of i2b2 plug-ins for query tools and data visualization.

The i2b2 data model is based on the “star schema” design where each row in the main database table represents a single “fact”. The facts are observations about a patient. Observations about a patient are recorded regarding a specific concept such as a lab value or medication order in the context of either an inpatient or outpatient encounter. This way of expressing a concept as an attribute in a row is known as the entity-attribute-value (EAV) model. It is very efficient to query data arranged in a star schema represented in an EAV format as a single index enables all patients’ data to be searched in one query. A screen shot of our i2b2 instance is listed in Figure 1.

The i2b2 platform is used by the NIH Clinical Translational Science Award (CTSA) network, and other academic health centers. i2b2 is funded as a cooperative agreement with the National Institutes of Health. The i2b2 platform will 1) enable data sharing across institutions; 2) construct extensible frameworks; 3) be able to utilize existing client and web interfaces and 4) make use of a controlled vocabulary.

We have access to a large-scale clinical data repository, MARS, which serves the University of Pittsburgh Medical Center (UPMC). MARS has been in existence since 1990 and contains over 300 million clinical records. MARS captures data at the point of care in both inpatient and outpatient settings. There are two major limitations to using this type of architecture for data-intense research projects, which we plan to overcome with the Megascope system. MARS is a text-based information system that imposes no structure on the report data as it enters MARS. While this makes it easy to accept new data, it presents an integration challenge when trying to harmonize data across reports. Second, there is no support for a query and analysis tool. Query access to the system for research purposes is limited to those who are familiar with the historical formats and content of various report types. However, through the use of the i2b2 platform and its support of controlled vocabularies and biomedical ontologies we can standardize the already collected data so we can use existing i2b2 query tools and reports.

We recognize that our research data needs to be stored and managed in an accessible architecture that can be used across institutions to enable data sharing. Megascope provides a platform which is flexible enough to support three major functions: 1) study cohort selection; 2) patient phenotype identification and 3) outcome assessment.
To date, our i2b2 system contains laboratory, demographics, pathology, medication (prescription) data and ICD9 diagnoses and procedure codes. We plan on adding the annotated data in the next few months as well as the genetic information. The Megascope application is displayed below:



GIANT

Progess Report on Technical Objective 2: Algorithm Development

Our IBD cohort is defined in three separate but overlapping groups. There are 1518 individuals who had at least one surgery for Crohn's Disease. Of these 1518, 262 have at least five years of follow-up and had their disease diagnosed within ten years of first being seen at our facility and have genetic information available. We consider this to be our gold set. An additional 332 have genetic information available and at least five years of follow-up but their disease was diagnosed outside of the ten-year window for initial diagnoses. The remaining patients (n=1017) will be also be used but their analysis may be limited.

Our Acute Pancreatitis (AP) cohort is being identified in a two-step process. We identified all unique AP (via ICD-9 code) admissions from 2005 through 2009 for patients having their 1st AP admission during the study period (n=2924). From this set, we have randomly selected 100 AP visits (50 at a community hospital and 50 at our main teaching hospital). The electronic charts of those 100 visits are in process of being reviewed by Drs. Yadav and Greer in our GIANT annotation tool. We will track the accuracy of the assignment of ICD-9 codes and assess if adding laboratory data confirms the presence of AP. Using our results, we will select the AP study cohort to be used for the algorithm development.

We are in the process of developing a methodology for identifying inconsistent data within each cohort. Since we are integrating data from many sources, we need to have a mechanism to identify data anomalies within a patient profile. We plan to spend the 1st quarter of FY13 on this task.

To support the natural language processing needed for our algorithm development, we built a GI clinical phenotyping pipeline (figure 3) within Megascope. To capture those data elements that are only identifiable by domain experts, we developed a web-based annotation tool, GIANT, to enable researchers to annotate de-identified clinical reports. The application design focuses on providing users with an intelligent workspace, by displaying annotation forms and de-identified reports with the same view, automatic report queuing and providing easy access to annotation guidelines and data definitions. The application produces user statistics to report agreement between multiple annotators who are reviewing the same report. Our tool was built using the Django (www.djangoproject.com) web framework, which is an open-source project built on the Python (www.python.org) programming language. Django provides a scalable platform for rapid web application development. The annotation tool features include controlled user access, database support, progress reporting, task-specific error checking and a site administration interface.

There are two output streams for GIANT. The first output is the report annotations completed by the clinical expert that will be imported into i2b2. The second output is the list of concepts identified in ODIE that appear most frequently in documents. This concept generator is used for feature selection to comprise the elements in the predictive model.
We worked with both the hospital and university information technology security teams to ensure that the system met the security policies. The system only houses de-identified data and is accessible from only the hospital or university network. Account management is done through our GIANT developer, Greg Gardner. Clinical users can only enter data and are not permitted any query or view access to the data. Each account also has a security question in addition to a password to access the system.

The GI surgical oncology group is also using GIANT for a project on pancreatic surgery outcomes. We think this extension of the application will help us improve usability as we further use in the tool in GI diseases.

Based on our experience with GIANT, we have completed 3 releases since November, 2011. We plan issue version 4.0 using the feedback from the AP portion of this project.

A central component of our pipeline uses our previous work developing the Ontology Development and Information Extraction (ODIE) system for ontology based annotation of clinical documents. The ODIE toolkit encompasses a suite of services for ontology-based text annotation (OA) and ontology enrichment (OE) combined with the ODIE workbench for user interaction, analysis, and visualization. Analysis engines for OA and OE, are executed in the Unstructured Information Management Architecture (UIMA) environment, an open-source, Apache-supported component software platform for unstructured information analysis.

Ontology development and enrichment

As we started examining the operative notes that were annotated in GIANT for Crohn's disease surgery, we recognized that the operative procedure names were complex. In processing the notes through ODIE, we could not find an ontology which recognized many of the operative procedure names. In discussing this issue with ontology domain experts, we realized that the GI surgery domain is not well represented in standard ontologies. So, we are adding each of the procedure terms to our Megascope ontology and will contribute this ontology to the National Center for Biomedical Ontology (www.bioontology.org) upon completion.

ODIE identifies both concepts (CUI) and semantic types (TUI) found in the narrative reports. We analyzed the output by positive finding via GIANT (Crohns/yes;Crohns Surgery/yes or Crohns/no;Crohns Surgery no) and manually reviewed the disagreements. Our initial screen did not identify many multi-word terms so we needed to adjust our configuration file to identify more multi-word phrases since most of our output was limited to a single word.

We have purchased and installed Research v7.6 software to facilitate the building of the Bayesian networks. We are currently defining the input requirements.

Progress Report on Technical Objective 3: Proof-of-Principle Study

We have completed an electronic review of the 594 patients who are in our NIDDK study. The electronic review was done via GIANT. We will next be adding their genetic information to our i2b2 instance as well as exporting their data to the Hugin application.

We plan on meeting with Dr. Yadav in the next month to explore using the SAPS (Severity of Pancreatitis Study) as the proof-of-principle study for the AP cohort.

In order to test our ability to work with routine EMR data for development of practical predictive models, we also analyzed an already-completed large EMR dataset obtained with an existing IRB-exempt permission for 3,925 diabetic patients at risk for nonalcoholic fatty liver disease (NAFLD). We found that seven simple indices calculated from routine clinical laboratory data along with BMI and age had strong predictive value over 5 years of observation for 1) the adverse outcomes of total deaths and liver-related deaths; 2) need for liver transplant; 3) hepatic encephalopathy; and 4) hepatocellular carcinoma. Logistic regression analysis confirmed that the clinical predictive value of the indices was independent of the presence or absence of hepatic steatosis on abdominal imaging.

Key Research Accomplishments

Abstracts presented at American Gastroenterology Association Digestive Disease Week 2012:

- Tomizawa Y, O'Connell MR, Saul MI, Slivka A, Whitcomb DC. Prevalence and Validation of Alcoholism Diagnosis in Chronic Pancreatitis (CP) Patients [abstract]. Gastroenterology. 2012 May; 142 (5):S-462.

- Raina A, Yadav D, Regueiro MD, Krasinskas AM, Saul M, Sapienza D, Binion DG, Hartman D. Mucosal IgG4 Cell Infiltration in Ulcerative Colitis (UC) is Linked to Disease Activity and Primary Sclerosing Cholangitis [abstract]. Gastroenterology. 2012 May; 142 (5):S-686.

- Dunn MA, Behari J, O'Connell M, Furlan A, Aghayev A, Gumus S, Saul,MI Bae KT. Noninvasive Hepatic Fibrosis Scores Predict Liver-Related Outcomes in Diabetic Patients [abstract]. Gastroenterology. 2012 May; 142 (5):S-1016.

- Gajendran M, Watson AE, Schraut WH, Regueiro M, Szigathy E, Dunn MA, Rivers CW, Binion DG. Patterns of Clinical Recurrence and Health Care Utilization in Post-Operative Crohn's Disease: Two Year Outcomes[abstract]. Gastroenterology. 2012 May; 142 (5) S-789.

Abstract presented at 2nd ACM SIGHIT symposium on International health information 2012:

- Sverchkov, Y.; Visweswaran, S.; Clermont, G.; Hauskrecht, M.; Cooper, G.F.; A multivariate probabilistic method for comparing two clinical datasets,Proceedings of the 2nd ACM SIGHIT symposium on International health informatics, 795-800 2012.

Papers published:

- Mehrotra A, Dellon ES, Schoen RE, Saul M, Bishehsari F, Farmer C, Harkema H. Applying a natural language processing tool to electronic health records to assess performance on colonoscopy quality measures. Gastrointestinal endoscopy. 2012 Apr 4. PMID: 22482913.

- Saligram S, Lo D, Saul M, Yadav D. Analyses of Hospital Administrative Data that Use Diagnosis Codes Overestimate the Cases of Acute Pancreatitis. Clinical gastroenterology and hepatology: the official clinical practice journal of the American Gastroenterological Association. 2012 Apr 10. PMID: 22504004.

Reportable Outcomes

- Coordinated Research meeting for Walter Reed and Pittsburgh co-investigators and staff in Pittsburgh on April 27, 2012 (agenda included in appendix).

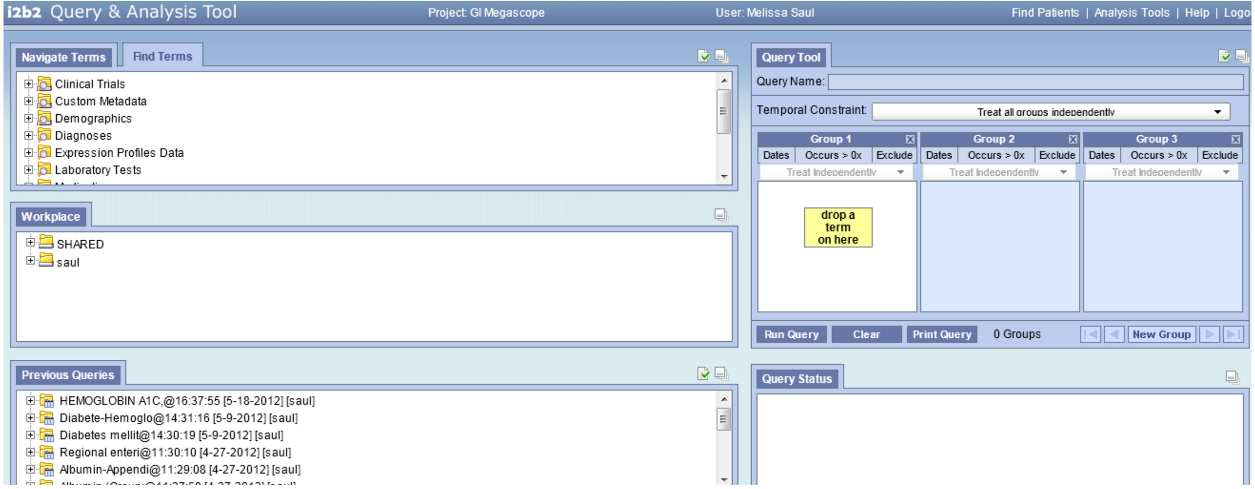
- Purchased and installed a Dell Poweredge R710 Server with 24GB of memory for use as a development machine for the Megascope application and processing of genetic data. The server resides at the University of Pittsburgh Network Operations Center (NOC).
- Melissa Saul and Michele Morris presented our work-in-progress to the University of Pittsburgh Medical Center (UPMC) Information Services Division (ISD) Interoperability Team on May 9, 2012

Conclusions

We have assembled a dynamic and strong team from both gastroenterology and informatics from the University of Pittsburgh and Walter Reed Army Medical Center. We have built our infrastructure according to plan and are ready to begin the collaborative algorithmic development task as stated in our proposal.

Figures

Figure 1 i2b2 Query Tool



Project: GI Megascope

User: Melissa Saul

- Clinical Trials
- Custom Metadata
- Demographics
- Diagnoses
- Expression Profiles Data
- Laboratory Tests

- SHARED
- saul

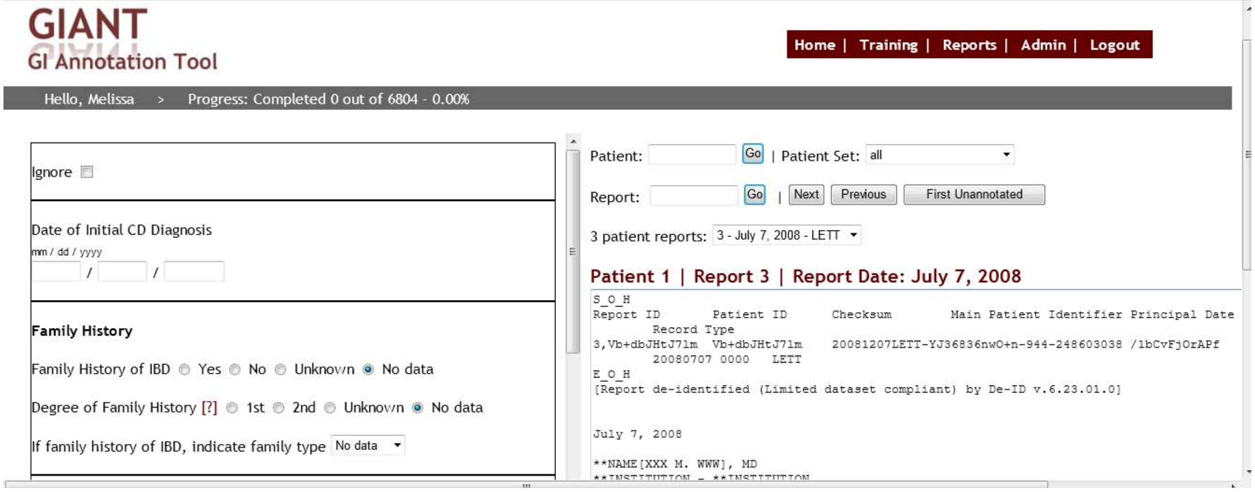
- HEMOGLOBIN A1C,@16:37:55 [5-18-2012] [saul]
- Diabete-Hemoglo@14:31:16 [5-9-2012] [saul]
- Diabetes mellit@14:30:19 [5-9-2012] [saul]
- Regional enteri@11:30:10 [4-27-2012] [saul]
- Albumin-Appendi@11:29:08 [4-27-2012] [saul]

Query Name:

Temporal Constraint: Treat all groups independently

0 Groups

Figure 2 – GIANT user interface



Hello, Melissa > Progress: Completed 0 out of 6804 - 0.00%

Ignore □

Date of Initial CD Diagnosis

mm / dd / yyyy

/ /

Family History

Family History of IBD ○ Yes ○ No ○ Unknown ● No data

Degree of Family History [?] ○ 1st ○ 2nd ○ Unknown ● No data

If family history of IBD, indicate family type No data

Patient: | Patient Set: all

Report: |

3 patient reports: 3 - July 7, 2008 - LETT

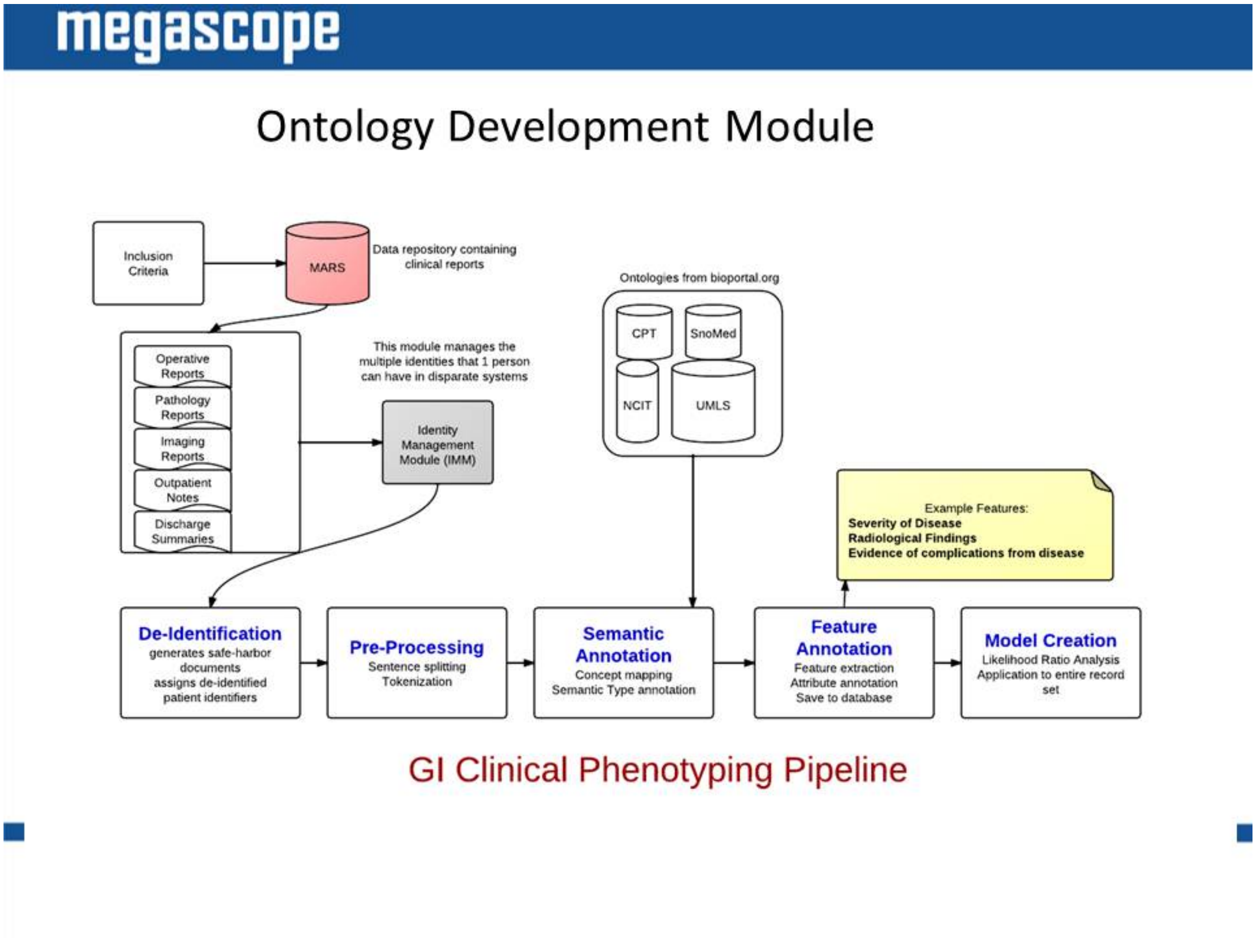
Patient 1 | Report 3 | Report Date: July 7, 2008

```
S_O_H
Report ID        Patient ID       Checksum        Main Patient Identifier Principal Date
        Record Type
3,Vb+dbJHtJ7lm  Vb+dbJHtJ7lm    20081207LETT-YJ36836nwO+n-944-248603038 /1bCvFjOrAPf
        20080707 0000   LETT
E_O_H
[Report de-identified (Limited dataset compliant) by De-ID v.6.23.01.0]


July 7, 2008

**NAME[XXX M. WWW], MD
**INSTITUTION - **INSTITUTION
```

Figure 3 – GI Clinical Phenotyping Pipeline



megascope

Ontology Development Module

GI Clinical Phenotyping Pipeline

Appendices

Appendix 1 – Agenda for April 27, 2012 Joint Meeting

Agenda

Framework for Smart Electronic Health Record Linked Predictive Models to Optimize Care for Complex Digestive Diseases Project

Hosted by: University of Pittsburgh Department of Biomedical Informatics (DBMI)

April 27, 2012

I. Overview of Project Goals and Study Aims

II. Patient Cohort Identification Methods
   a. Crohn's Disease (CD)
   b. Acute Pancreatitis (AP)

III. Work-in-Progress Review of DBMI Software Tools
   a. GIANT
   b. ODIE- http://www.bioontology.org/ODIE
   c. I2b2 - https://www.i2b2.org/software/index.html
   d. Medical Archival Retrieval System (MARS)

IV. Data Collection Process
   a. Progress to date for both CD and AP cohorts
   b. Discussion on final output needed for Bayesian Network Application

V. Bayesian Network Development and Application

VI. Next Steps

Meeting Information

Location: Department of Biomedical Informatics
M-183 Parkvale Building
200 Meyran Avenue
Pittsburgh, PA 15261

http://maps.google.com/maps?q=%22200+meyran+ave%22++pittsburgh+PA++15260&hl=en&om=1&hnear=200+Meyran+Ave,+Pittsburgh,+Pennsylvania+15213&t=h&z=15

Parking: Public Parking is available next to building or across the street from building.

University of Pittsburgh Attendees

Michael Dunn, MD
Julia Greer, MD
Shyam Viswaswaram, MD, PhD
Melissa Saul
Michele Morris